5-25-2018

Assessing Subjectivity in Environmental Sensor Data Post Processing via a Controlled Experiment

Amber Spackman Jones
Utah State University

Jeffery S. Horsburgh
Utah State University

David P. Eiriksson
University of Utah

Recommended Citation
Jones, Amber Spackman; Horsburgh, Jeffery S.; and Eiriksson, David P., "Assessing Subjectivity in Environmental Sensor Data Post Processing via a Controlled Experiment" (2018). *Publications.* Paper 149.
https://digitalcommons.usu.edu/water_pubs/149

Ecological Informatics 46 (2018) 86–96

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Assessing subjectivity in environmental sensor data post processing via a controlled experiment

Amber Spackman Jones[a,*], Jeffery S. Horsburgh[b], David P. Eiriksson[c]

[a] Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, United States
[b] Department of Civil and Environmental Engineering, Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, United States
[c] Global Change and Sustainability Center, University of Utah, Salt Lake City, UT 84112, United States

ARTICLE INFO

Keywords:
Sensor data
Quality control
Water quality
Data correction
Monitoring

ABSTRACT

Collection of high resolution, in situ data using environmental sensors is common in hydrology and other environmental science domains. Sensors are subject to drift, fouling, and other factors that can affect the quality of the measurements and their subsequent use for scientific analyses. The process by which sensor data are reviewed to verify validity often requires making edits in post processing to generate approved datasets. This quality control process involves decisions by technicians, data managers, or data users on how to handle problematic data. In this study, an experiment was designed and conducted where multiple participants performed quality control post processing on the same datasets using consistent guidelines and tools to assess the effect of individual technician on the resulting datasets. The effect of technician experience and training was also assessed by conducting the same procedures with a group of novices unfamiliar with the data and compared results to those generated by a group of experienced technicians. Results showed greater variability between outcomes for experienced participants, which we attribute to novice participants' reluctance to implement unfamiliar procedures that change data. The greatest variability between participants' results was associated with calibration events for which users selected different methods and values by which to shift results. These corrections resulted in variability exceeding the range of manufacturer-reported sensor accuracy. To reduce quality control subjectivity and variability, we recommend that monitoring networks establish detailed quality control guidelines and consider a collaborative approach to quality control in which multiple technicians evaluate datasets prior to publication.

1. Introduction

Collection of high resolution, in situ data using environmental sensors is common in hydrology and many other environmental science domains (Hart and Martinez, 2006; Pellerin et al., 2016; Rode et al., 2016). Sensors are subject to drift, fouling, and other factors that can affect the quality of the measurements and their subsequent use for scientific analyses, particularly when sensors are deployed in aqueous or other harsh environments where scaling, biological growth, or other adverse conditions can occur (Campbell et al., 2013; Pastorello et al., 2014; Wagner et al., 2006). Sensor datasets are typically subjected to quality control (QC) post processing procedures to verify their validity prior to use in scientific analyses. However, documentation of QC procedures used by scientists in journal publications based on the data often do not contain sufficient detail to permit reproducibility, and it is also rare for both raw and quality controlled data to be shared so that subsequent users can examine the degree to which the data had been modified prior to analysis (Daly et al., 2005). The overall level of uncertainty in observations made using in situ sensors is dependent not only on the accuracy and precision of the sensor, but also on the sensor deployment technique, environmental conditions, and the subsequent procedures used to post-process the data (Gries et al., 2014; Wagner et al., 2006).

Several studies have investigated and explored automated procedures for detecting anomalies and problems in sensor datasets (Dereszynski and Dietterich, 2007; Fiebrich et al., 2010; Hill et al., 2009; Meek and Hatfield, 1994; Moatar et al., 2001; Shafer et al., 2000; Sheldon, 2008; Taylor and Loescher, 2013; White et al., 2010). Once identified, options for dealing with problematic data include removing data from a time series of observations, retaining data with annotations, setting the values of problematic observations to a “NoData” value, or altering the data values based on algorithms that use adjacent data values or patterns in data at other locations or of other variables (Campbell et al., 2013; Horsburgh et al., 2015). These algorithms are

[*] Corresponding author.
E-mail addresses: amber.jones@usu.edu (A.S. Jones), jeff.horsburgh@usu.edu (J.S. Horsburgh), dave.eiriksson@utah.edu (D.P. Eiriksson).

https://doi.org/10.1016/j.ecoinf.2018.05.001
Received 3 April 2018; Received in revised form 18 May 2018; Accepted 20 May 2018
Available online 25 May 2018

typically good for plausibility checks of meteorological data, but are less commonly implemented in water quality monitoring applications where QC is more often a subjective process that requires judgement from a technician whose job it is to post process the data (Daly et al., 2005; ESIP EnviroSensing Cluster, 2014; Qu et al., 2016). Some post processing can be automated, but for many scientists, research groups, and technicians, the QC process is manual and time consuming, so much so that it is not uncommon for scientists to hire undergraduate students or other, less-experienced personnel to complete this work.

Ideally, given the same dataset and the same QC guidelines, multiple data QC technicians would make the same decisions in data post processing. However, despite the development and implementation of guidelines aimed to promote consistency (e.g., Jones et al., 2017), we have faced ambiguity when performing post processing because it is not always obvious which correction procedures should be applied. We have also noticed inconsistencies between individuals performing QC post processing. Technicians with the same level of training, using the same input datasets and field notes, and using the same software tools may produce different results. Furthermore, we have observed inconsistency in results produced by technicians that do not have the same level of training or experience in field work, sensor deployment and maintenance, data collection, and QC post processing. It is clear that subjectivity in sensor data post-processing affects the overall quality and comparability of finished data products, but the degree to which this is the case has not been well described or quantified.

This study had two objectives. The first was to identify and quantify the degree of subjectivity among technicians performing QC on environmental sensor datasets to better understand how results are affected as multiple technicians participate in the QC process. The second was to assess differences in the outcome of the QC procedure for novice versus experienced technicians. To accomplish these objectives, an experiment was designed where multiple participants performed QC post processing on the same sensor datasets using a consistent set of guidelines and software tools. The effect of technician experience and training was assessed by conducting the same experiment with a group of novices who were unfamiliar with the data and who had never performed QC on environmental sensor data. Results from the novice group were compared to those generated by the more experienced technicians to quantify the impacts of individual technician and technician experience and report the observed degree of subjectivity in sensor data post processing.

Our familiarity with performing and observing QC post processing results from experience with high frequency environmental data collected by sensors in an ecohydrologic observatory monitoring Gradients Along Mountain to Urban Transitions (GAMUT). This monitoring network is part of Utah's iUTAH (innovative Urban Transitions and Arid-region Hydrosustainability) project, a state-wide, multi-institutional, multi-disciplinary effort. The sites, sensors, and methods used to design and operate the GAMUT network, including detailed descriptions of QC procedures, are documented by Jones et al., 2017.

2. Methods

2.1. Participant groups

Participants in this study were recruited to comprise two groups: novices unfamiliar with QC and more experienced practitioners. We sought to compare the two groups with the anticipation that experienced users' processed results would converge toward a central tendency and that novice users' processed results would be more outlying. The novice group (n = 15) primarily consisted of undergraduate students participating in a summer undergraduate research experience. Additional novices included mentors for the students within this group. None of the novices had performed QC on environmental sensor data prior to participating in the study and were not familiar with the QC process, although several of them had participated in field data collection activities.

Table 1
Mean values of participants' responses to a survey regarding prior experience and difficulty of this exercise. Responses are on a 1 (low) – 10 (high) scale.

Group	Computing experience	Familiarity with water quality data	Familiarity with field work	Difficulty of exercise
Novice (n = 15)	4.00	5.14	6.43	5.86
Experienced (n = 13)	8.23	8.62	9.00	5.00

The experienced group (n = 13) consisted of participants who work with environmental sensor data and perform QC as part of their full-time job or as part of their active research. These participants included full-time watershed field technicians, data managers, faculty, undergraduate students, and graduate students, all of whom regularly work with time-series sensor data and perform QC tasks as part of their research work. All experienced participants had received formal and/or informal training in performing QC on environmental sensor datasets, although this training was not part of this study and was not standardized across the participants. So, while we were able to recruit a number of experienced participants, we were not able to standardize the level of experience or training received by those in the experienced group. According to self-reported experience relevant to this exercise (Table 1), experienced participants were more proficient with computing as well as more familiar with water quality data and fieldwork than novice participants. Additional results and details of this survey are discussed in subsequent sections (2.1, 3.1, 3.5).

2.2. Selected datasets

Three raw, continuously and simultaneously measured sensor datasets were selected for this experiment from an in situ water quality monitoring site in the Logan River at Mendon Road, near Logan, UT. Measured variables included water temperature, specific conductance, and pH recorded every 15 min using a YSI/Xylem EXO2 multiparameter water quality sonde (https://www.exowater.com/exo2) equipped with a central wiper. Table 2 lists the specifications for the sensors used to observe these three variables. These variables were chosen because they are commonly sensed at aquatic monitoring sites and because they represent relatively well-known environmental phenomena understood by both expert and novice participants. The duration of the input data was limited to a period of one year (January 1, 2014 to December 31, 2014) to ensure that there would be significant and varied quality-related issues with the data that needed to be corrected and to balance the time required for participants to complete the study.

During the period over which the data were collected, field maintenance was performed at the selected site on a monthly or bi-weekly basis, depending on seasonal water quality. In general, regular maintenance site visits consisted of removing the water quality sonde to observe its condition and clean the sensors if necessary, verifying that the integrated wiper was functional, field calibration checks using standard and traceable reference solutions, and recalibration of the sensors as necessary. These procedures are typical of the type of maintenance performed at aquatic monitoring sites. While these field visits were part of a larger quality assurance plan designed to maximize the quality of the raw data collected in the field, there were still issues with the raw data that needed to be corrected in post processing.

A record of field maintenance, other activities, and notes related to the selected site and variables was compiled over the year of interest. Each participant was provided a copy of this record as a reference for performing QC. This type of field record is essential when conducting QC post processing because the selection of points for and the choice of post processing edits is directly related to details about when site visits were conducted, when sensors were out of the water, when calibrations

Table 2
Specifications for the sensors measuring the variables used for this study (Xylem, 2012). Effective accuracy refers to the values applicable to the data used in this study.

Variable	Sensor Model	Range	Sensor Accuracy	Effective Accuracy
Water temperature	YSI EXO 599870–01		−5 to 35 °C: ± 0.01 °C; 35 to 50 °C: ± 0.05 °C	± 0.01 °C
Specific conductance	YSI EXO 599870–01	0 to 200 mS/cm	0–100 mS/cm: ± 0.5% of reading or 0.001 mS/cm, whichever is greater; 100–200 mS/cm: ± 1% of reading	± 2 μS/cm
pH	YSI EXO 599795–02	0 to 14 pH units	± 0.1 pH units within ± 10 °C of calibration temperature; ± 0.2 pH units for entire temperature range	± 0.1 pH units

or other maintenance actions were performed, and observations made by technicians in the field. A copy of that record is shared with the study dataset in the HydroShare system (Jones et al., 2018).

None of the participants were given information about the location of the monitoring site at which the raw data were collected to ensure that results were not biased by participants' prior knowledge of (or lack of knowledge of) conditions at that site. The raw data used as the basis of this study are published as part of the overall study dataset (Jones et al., 2018).

2.3. Quality control process

Each participant was asked to perform QC post processing on the three separate time series in the selected dataset using the Observations Data Model (ODM) Tools software (Horsburgh et al., 2015). ODM Tools was developed in Python as a graphical user interface for visualizing and performing scripted quality control post processing of environmental sensor time series data. The software interacts with a relational database instance of the Observations Data Model (Horsburgh et al., 2008) within which the data were stored.

ODM Tools was selected as the software for this experiment because it provides a graphical user interface and relatively straightforward tools for performing the most common types of edits needed for post processing the types of data we chose for this experiment. Users can interactively select data to be edited on a plot or by using custom date or value filters and then click buttons on the toolbar to shift data, interpolate values, perform a drift correction, etc. A benefit of ODM Tools is that it automatically records all data selections and edits to a Python script. All participants saved and submitted their Python scripts, thus creating a complete record of the changes made by each participant as well as any comments they inserted into the Python script to annotate their QC choices. ODM Tools is open source (https://github.com/ODM2/ODMToolsPython) and can easily be installed on many different computers.

Participants in the novice group completed the experiment during a single day in a computer lab on Utah State University's campus. Novices met as a group and received a brief orientation on water quality, aquatic sensors and data, and environmental conditions that may affect observations via an oral presentation. Novices then received a 30-minute demo of the functionality of ODM Tools to ensure that they could operate the software to complete the experiment. Participants in the novice group then conducted their post-processing in the lab. Several instructors were present to answer participants' questions about how to operate the software but did not address questions about whether corrections should be made, which corrections should be made, or the extent to which a correction should apply in efforts to eliminate any bias introduced by the instructors. Given that the experiment was conducted with the novice group in a single session, these participants were limited in the time they could spend on QC (approximately 3 h), and most did not complete QC on all three time series.

Experienced participants were provided with the information needed to connect to their own ODM database containing the study data, and they completed the experiment independently on their own time, on their own computers, in their own offices. Each of the experienced participants had already been trained in the use of the ODM Tools software and had used the software extensively prior to the experiment. Similar to the novice group, we did not answer questions that may have introduced bias into experienced participants' decisions about which corrections to implement. We did not control whether they worked in single or multiple sessions. All experienced participants performed QC to completion on all three time series. Experienced participants reported spending 3–10 h on the exercise, with a median duration of 4 h.

2.4. Exit survey

As a final step, participants were asked to complete an exit survey, which was used to elicit information about their level of experience related to computing and water quality data and field work (Table 1) as well as their reactions to the exercise. This survey was developed using the Qualtrics software, and participants completed it online. The results of the survey can be explored at http://data.iutahepscor.org/surveys/survey/QCEXP# and are also published as part of the dataset related to this work (Jones et al., 2018).

2.5. Data management and analysis

To facilitate the experiment, multiple, replicate ODM databases containing the raw observational data were created. Each participant was assigned an individual ODM database, to which they connected, completed post processing edits independently, and saved their processed datasets. All of the processed data were subsequently collated into a single ODM database to facilitate simpler queries and data access. To simplify analysis, the data were exported into multiple comma-separated text files, where each file contains all of the post-processed time series for a single variable (e.g., temperature). Rows in these files represent the date and time of each observation, and each column corresponds to a single participant with the post processed data contained in the table. This form of the data was used for the bulk of the visualization and analysis reported below. All visualization and analyses were performed in Matlab, and associated scripts are contained with the published dataset (Jones et al., 2018).

Although the link between the final, post processed datasets and the Python scripts created by the participants was maintained, the data were anonymized to protect the identities of the study participants. Each participant was assigned an arbitrary integer identifier at the outset of the study. These identifiers are used in the final, anonymized versions of the data used in this study, which are published in the HydroShare data repository (Jones et al., 2018).

3. Results and discussion

Figs. 1-3 (panels a, d) show the results of QC performed by all participants for temperature, pH, and specific conductance, respectively. There is overlap between participants' results and the raw data, and the full time scale makes it difficult to distinguish individual

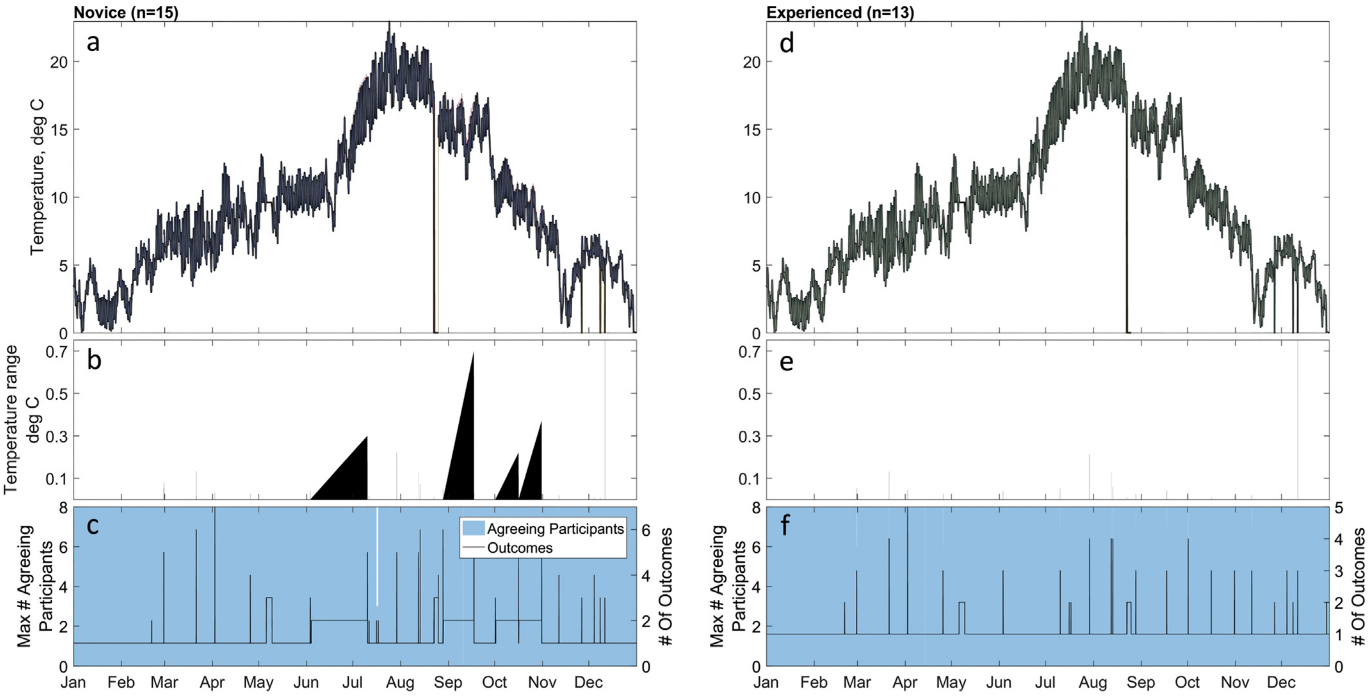

Fig. 1. Results for water temperature for novice participants (a-c) and experienced participants (d-f). Processed results are shown by colored lines and the raw data as a thicker black line (a, d). The range is the difference between the maximum and minimum values between participants at each time stamp (b, e). To determine the total number of outcome values among participants and the maximum number of agreeing participants (c, f), results were rounded to 0.01° C, and equivalent results between participants were binned at each time stamp.

results. Some excursions from the raw data are obvious as participants made QC choices that deviated from the majority, but in most cases, individual participants' responses obscure each other, an important result indicating a high degree of agreement among them. Section 3.2 focuses on and illustrates periods for which there is greater deviation between outcomes. To provide greater insight into results, these figures also indicate the full range of outcomes as the difference between maximum and minimum of all participants' results at each time stamp (panels b, e), as well as the maximum number of agreeing participants and corresponding number of outcomes at each time stamp (c, f). Results from these plots for experienced participants are summarized as averages and proportions of the data in Table 3. As shown in Fig. 4, the range of values for the average amount by which each participant changed data from the original, raw data to the quality controlled

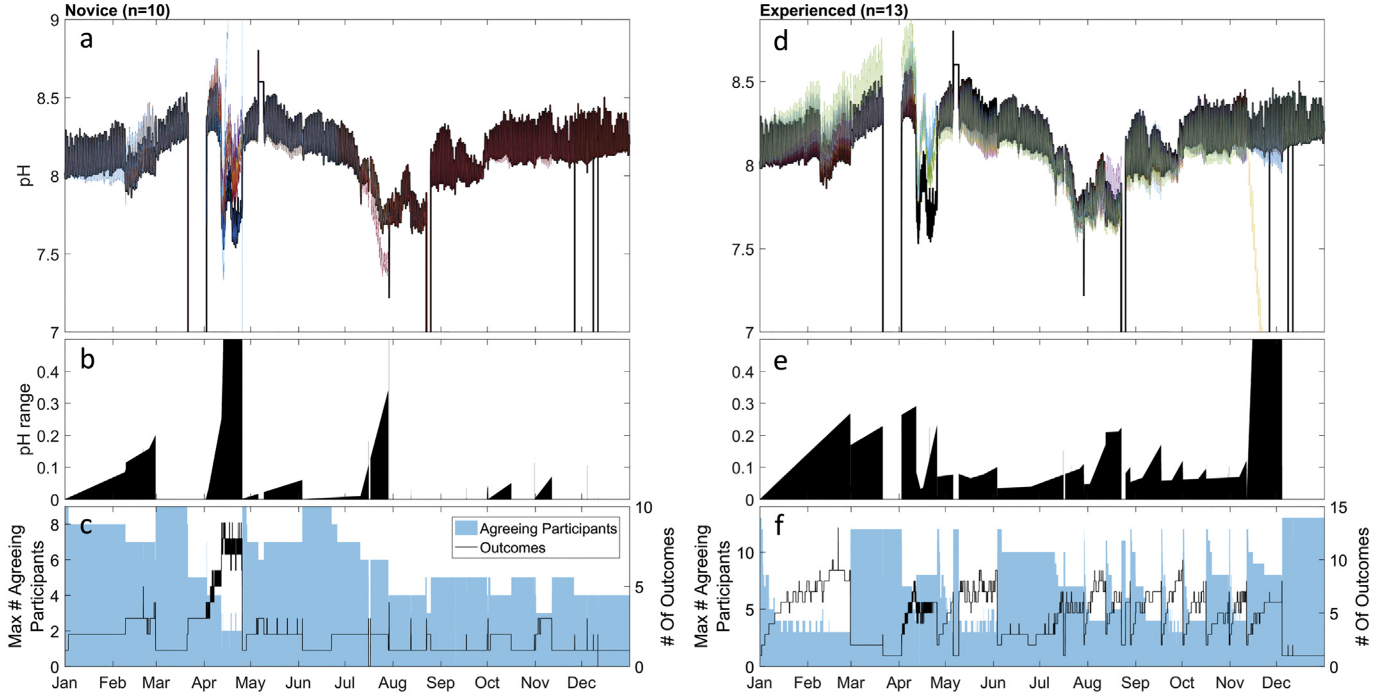

Fig. 2. Results for all participants for pH. See Fig. 1 for panel descriptions. The range scale (b, e) cuts off large values resulting from errors in drift correction. To determine the maximum number of agreeing participants and total number of outcomes at each time stamp (c, f), results were rounded to 0.01 pH unit.

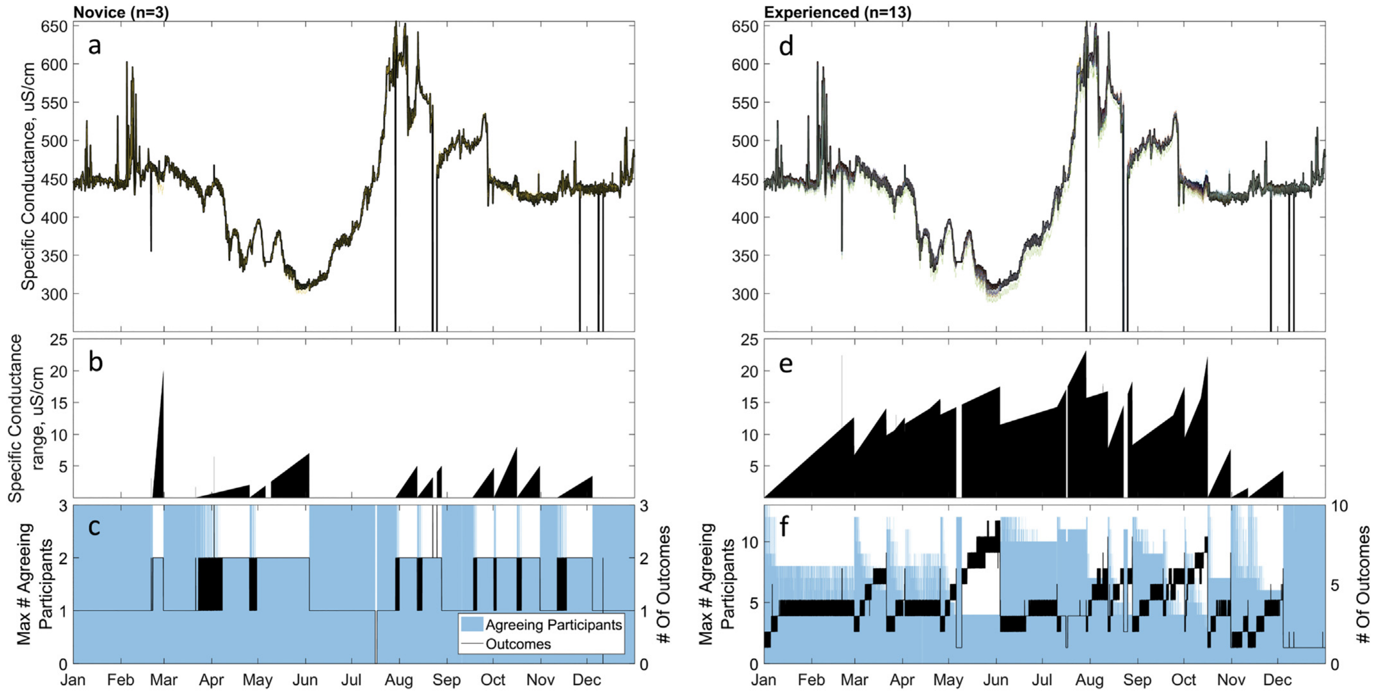

Fig. 3. Results for all participants for specific conductance. See Fig. 1 for panel descriptions. To determine the maximum number of agreeing participants and total number of outcomes at each time stamp (c, f), results were rounded to 1 μS/cm.

Table 3
Summary of agreement for experienced participants. To determine agreement and number of outcomes, results were rounded to 0.01° C, 0.01 pH units, and 1 μS/cm.

Metric	Temperature	pH	SpCond
Average of maximum number of agreeing participants (n = 13)[a]	12.94	7.43	8.40
Average number of outcomes (n = 13)[b]	1.03	4.68	3.51
Percent of data where all participants agreed[c]	97.5%	12.4%	11.5%
Percent of data where the range of values was within sensor accuracy[d]	99.4%	60.9%	17.9%

[a] Determined by calculating the highest number of agreeing participants at each time stamp and averaging over the entire record.

[b] Determined by finding the total number of outcomes at each time stamp and averaging over the entire record.

[c] Determined as the proportion of data in which all participants agreed on outcomes.

[d] Determined by comparing the range of participant responses with the manufacturer reported sensor accuracy (Table 2) at each time stamp.

versions depended on variable and experience level.

As demonstrated by these plots and metrics, variation in the post processed datasets from both participant groups in our experiment was observed. There are nuances in the variability that relate to the participant group (Section 3.1), the variable and associated QC practices (Section 3.2), the relative significance compared to sensor accuracy (Section 3.3), and ranges of summary statistics of the datasets for experienced participants (Section 3.4). We also describe results related to participants' handling of problematic data values (Section 3.5), annotations in QC scripts (Section 3.6), and errors in the QC process (Section 3.7).

3.1. *Variation between participant groups*

Contrary to our original research hypothesis, both overall variability in the processed results (Figs. 1-3) and the degree of difference between processed results and the raw data (Fig. 4b) were found to be greater for the experienced group than the novice group. Based on observations and discussions with participants, we conclude that participants in the experienced group were more willing to alter raw data versus those in the novice group. During the session with novice participants, we noticed hesitancy to make changes to the data. As altering data is part of typical QC practice, experienced participants were more comfortable with making changes. We also found that drift corrections were performed by all of the experienced participants, but only by a few of the novice participants, indicating that novices were less willing to make this kind of correction. This may reflect some degree of misunderstanding of the drift correction procedure among novices or that their perception of the need for drift correction is different than that of the experienced group (see Section 3.6 for an example). This assessment of novice versus experienced participants' processed data is limited given that most novice participants did not complete QC on all variables (n = 15 for temperature, n = 10 for pH, n = 3 for specific conductance), and the results subsequently described focus on the outcomes of the experienced participants.

The different attitudes between novice and experienced participants is exhibited by results of one of the questions in the exit survey (Fig. 5). When asked to identify the most challenging aspect(s) of the exercise, a majority of experienced participants indicated that decision making aspects presented the greatest barrier. Novice users, on the other hand, more commonly indicated that unfamiliarity with the data and the process was a primary challenge. However, experienced users did not rate the overall difficultly of the exercise less than did novice users (Table 1). Users familiar with QC recognize the challenge of making decisions in altering datasets – underscoring our observation of the subjective nature of the QC process.

3.2. *Differences between variables: drift correction*

There was greater agreement within and across both participant groups for variables that do not undergo regular calibration in the field and, therefore, do not need drift correction in post processing (e.g., temperature – Fig. 1, Table 3, and Fig. 4a). For temperature, all

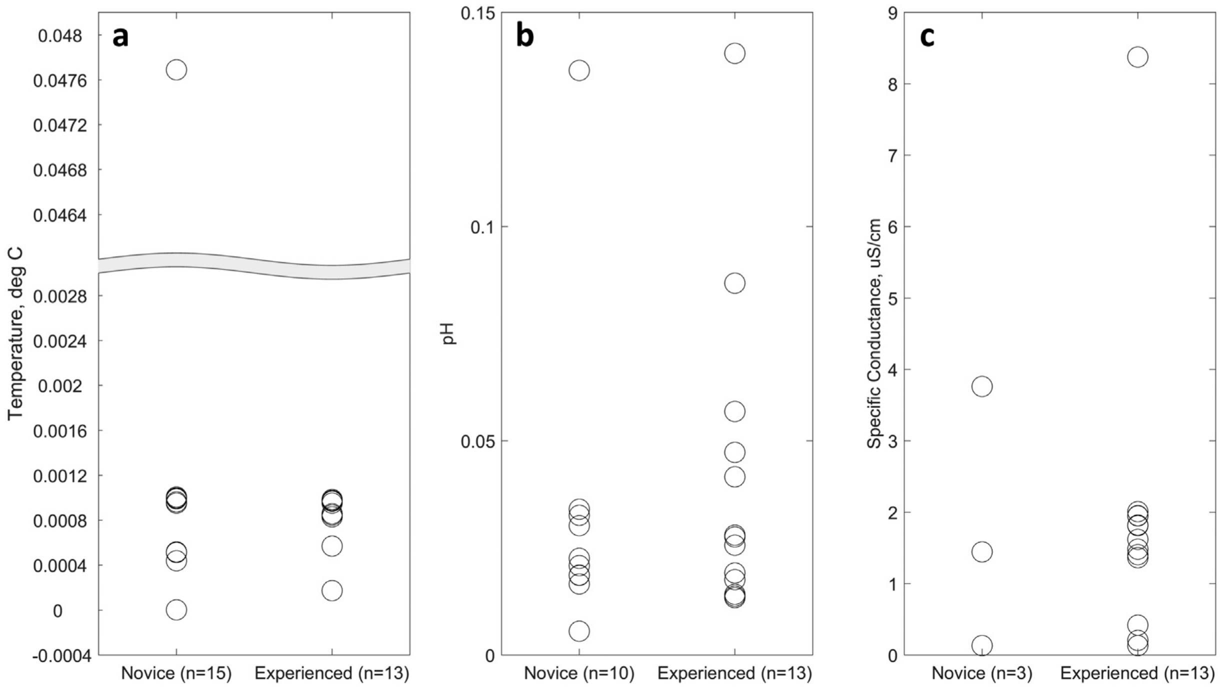

Fig. 4. Mean deviation from raw data for each participant organized by variable and experience level for (a) temperature, (b) pH, and (c) specific conductance. Values were determined as the mean of the differences between raw and processed data at each stamp for each participant. Note that data that were set to the "no data" value of −9999 were excluded from this analysis.

experienced participants agreed on 97.5% of the data, while for pH and specific conductance, all experienced participants only agreed 11–13% of time. This result was anticipated and underscores the importance of technological measures such as automated wipers and anti-fouling sensor coatings in extending the deployments of sensors affected by biological growth and fouling. These preventative measures reduce the impact of fouling on the measurements themselves, but also reduce the uncertainty and subjectivity introduced during the quality control process if drift corrections can be avoided.

Consistency was observed in the post processed data from all experienced participants shortly after field visits and calibration events when technicians were able to reference fixed calibration points (i.e., the values from all participants were nearly the same after each calibration). However, as the time after a calibration event increased, so did the range of post processed data from different technicians (Figs. 2 and 3 – panels e and f). This is a product of: 1) the choice by a technician to perform a drift correction to close an offset in the raw data caused by instrument drift and/or fouling, and 2) the technician's choice of the offset value to use in the drift correction. A summary of drift correction offsets implemented by experienced participants is shown in Fig. 6. For any given calibration, approximately half of participants opted to perform a drift correction. Offsets ranged as high as 18.9 μS/cm for specific conductance and 0.19 for pH. This is further illustrated by the ranges in results (Figs. 2e and 3e) and the corresponding disagreement in processed datasets (Figs. 2f and 3f).

To understand the source of this variability, it is important to be familiar with the underlying process. A linear drift correction moves the points prior to a calibration up or down by a specified offset and regressively applies the correction to past values up to a selected point in time, which typically corresponds to the previous cleaning or calibration (Horsburgh et al., 2015; Wagner et al., 2006). The most recent point is shifted by the offset, the point associated with the previous

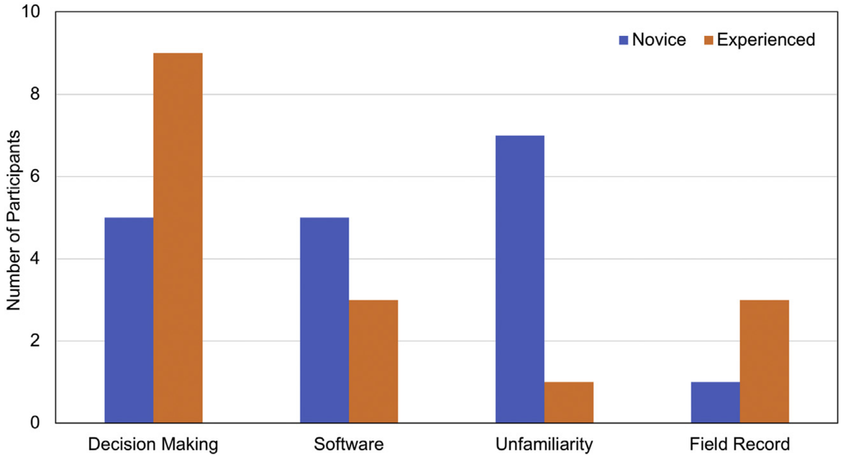

Fig. 5. Responses of participants to the question "What aspect of the exercise did you find the most challenging?" organized by experience level. Note that participants' responses could include multiple aspects.

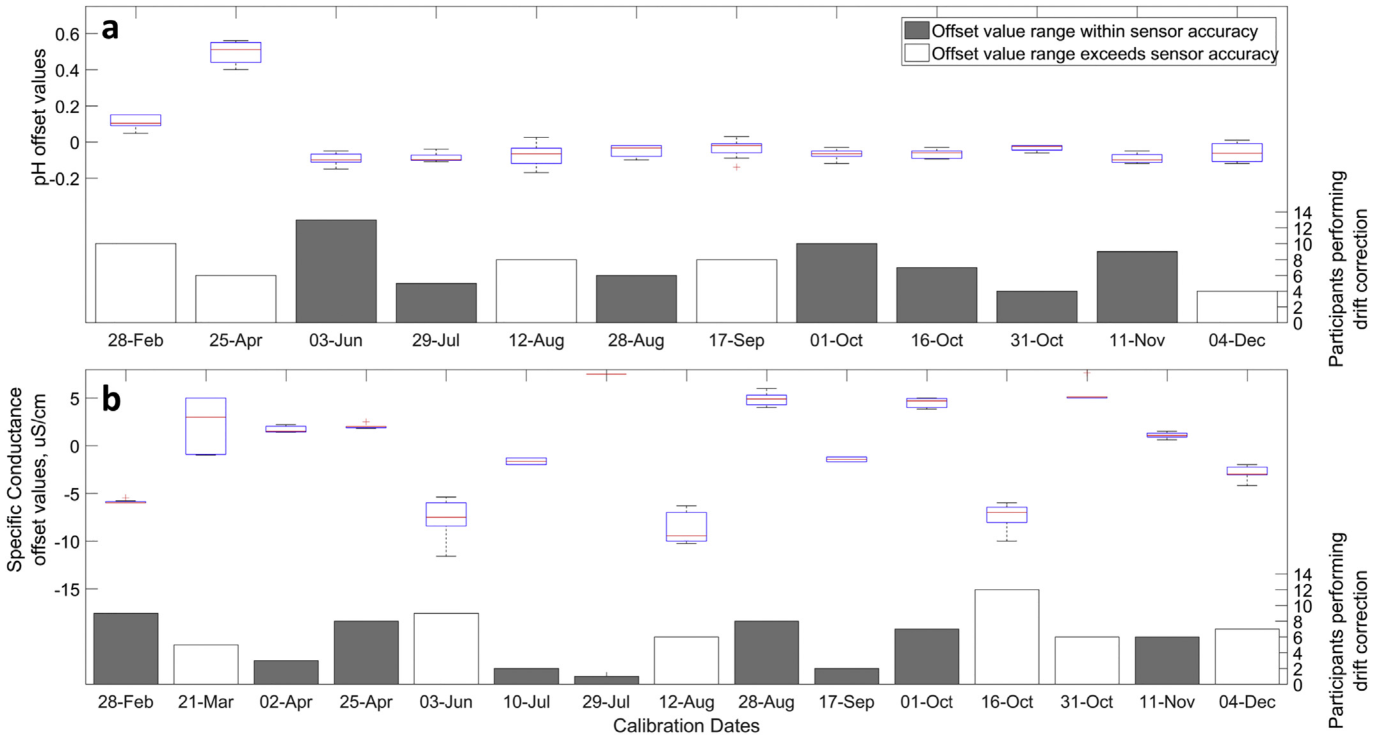

Fig. 6. Drift correction offsets selected by experienced participants for pH (a) and specific conductance (b). Bars represent the number of participants opting to drift correct (out of a total of 13). Box and whisker plots represent the range of offset values selected. Shaded bars indicate calibration events for which the range of participants' correction values were within manufacturer-reported sensor accuracy (0.01 for pH and 2 μS/cm for specific conductance).

calibration is not shifted at all, and each point in between is shifted proportional to the time distance between calibration events. When it is determined that a drift correction is to be performed, the technician chooses the exact point on which the shift is made and selects an offset. The values of offsets (the corrections illustrated in Fig. 6) may be estimated based on visual examination of the data, by calculating a value based on the slope of the data points before and after calibration, or by determining the error in the sensor reading observed in the calibration check based on the difference between the pre and post calibration readings. Because the field notes obtained for this exercise did not report the pre and post calibration readings, and because it is the easiest option, we conclude that most participants determined drift correction offsets by visually assessing how the corrected data should appear relative to post-calibration data.

Fig. 7 shows examples of two drift corrections performed by experienced participants. In these cases, most participants selected the same point to initiate drift correction, but the offsets varied. Fig. 7a corresponds to the pH calibration on October 1, 2014 for which 10 out of 13 participants opted to make a drift correction, and Fig. 7b corresponds to the specific conductance calibration on June 3, 2014 for which 9 out of 13 participants opted to make a drift correction. In examining participants' scripts for these two cases, those that did not perform a drift correction determined that it was more appropriate to: 1) interpolate erroneous data associated with the calibration and the period that the sensor was out of the water, 2) set the same data to −9999, or 3) leave the data unaltered. In all of these cases, a flag would be applied to annotate the data with a descriptive qualifier. This inconsistency in decision-making was observed for nearly all calibration events in this study. Out of 28 calibration events (13 for pH, 15 for specific conductance), only once did all participants decide to drift correct (pH on June 3, 2014) and only once did all participants opt for another method of data correction like linear interpolation (pH on July 10, 2014).

For both cases shown in Fig. 7, participants who elected to drift correct agreed that the data prior to the calibration event should be shifted down, presumably based on the pattern of data following the calibration event, but there was not agreement on the degree of the shift. In these particular examples, the selected offsets varied between −0.068 to −0.03 for pH and − 7.62 to −5.4 μS/cm for specific conductance. These discrepancies could be due to individual assessment of which point following a calibration is the "true" calibrated reading. When it is returned to ambient water, a sensor requires time to equilibrate, and this can result in some spurious points following field maintenance and calibration. The technician performing QC must make a determination of the first valid point post-calibration, which then influences the selection of the offset for the drift correction on the preceding data. These examples are representative of other calibration and drift correction occurrences in that there is broad disagreement between participants (see Fig. 6 for comparison to other calibration events). Rather than gravitating to a few consistent offsets (e.g., a bimodal distribution of results), or converging on a central tendency (e.g., a normal distribution), in these examples, participants' selected offsets are evenly distributed over the range of results (e.g., a uniform distribution). Similar patterns of distribution for other events were observed (data not shown).

3.3. Comparison to sensor accuracy

To put the scale of participants' processed results into perspective, the observed variability was compared to the manufacturer's reported sensor accuracy (Table 2) to assess whether the discrepancies are within these bounds. The manufacturer's reported accuracy was used as a conservative benchmark given that uncertainty of the field-measured values is likely higher than the accuracies reported by the manufacturer, which are determined in the laboratory under optimal conditions (e.g., Thoma et al. (2012) report lower instrument variability for temperature, pH, and specific conductance in laboratory than in field tests, and for these variables, the United States Geological Survey (USGS) criteria for "excellent" accuracy rating exceed the manufacturer-reported accuracies (Wagner et al., 2006)). Furthermore,

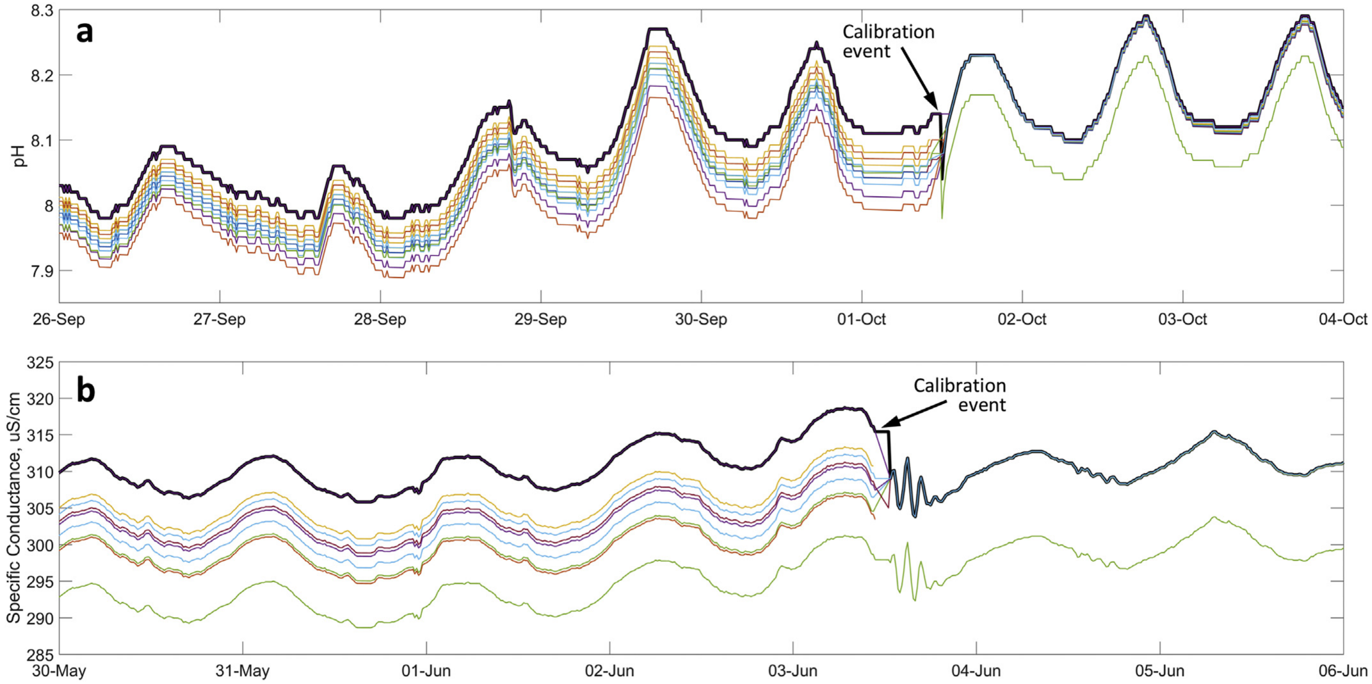

Fig. 7. Examples of calibration and drift correction by experienced participants for (a) pH and (b) specific conductance. Colored lines represent the QC results for individual participants, and the raw data is represented by a thick black line.

Table 4
Ranges of summary statistics for experienced participants.

Statistic	Temperature	pH	SpCond
Minimum	− 0.04 – −0.04 (0.00)	5.09[a]–7.63 (2.53)	288.1[a]–305.0 (16.9)
25th Percentile	5.33–5.43 (0.10)	8.03–8.07 (0.04)	410.2–421.3 (11.1)
Median	8.60–8.64 (0.04)	8.14–8.16 (0.02)	437.2–441.2 (4.0)
Mean	9.23–9.34 (0.11)	8.05–8.15 (0.10)	430.4–438.8 (8.4)
75th Percentile	12.58–12.81 (0.23)	8.25–8.28 (0.03)	454.8–462.7 (7.9)
Maximum	22.94–22.94 (0)	8.59–8.87 (0.28)	639.8–655.5 (15.7)

[a] A single participant's results included values of 0, which was clearly an erroneous outlier, so it was excluded from the range determination.

actual uncertainties may vary (e.g., over time) and may be dependent on conditions, whereas the manufacturer's reported accuracy is a standardized metric.

For temperature, the average deviation from raw data by experienced participants was well within the ± 0.01° C range (Fig. 4a), and the range of outcomes between participants was within ± 0.01° C for 99.4% of all data values (Table 3). For pH, the average deviation from raw data was within the sensor accuracy of ± 0.1 (Fig. 4b) for all experienced participants. The ranges of pH drift correction offsets selected by experienced participants (Fig. 6) were greater than ± 0.1 for 4 out of 13 calibrations, resulting in a range of outcomes within ± 0.1 for 60.9% of all data values (Table 3). We observed a similar pattern for specific conductance, though the range of processed results more commonly exceeded reported sensor accuracy, which we determined to be ± 2 μS/cm based on the range of data used in this study. The average deviation from the raw data (Fig. 4c) was within sensor accuracy. However, as shown in Fig. 6, the differences between experienced participants' results were outside of the range of sensor accuracy for 6 out of 15 calibrations, and these changes propagated through the data so that the range of outcomes was within 2 μS/cm for only 17.9% of the data (Table 3).

To summarize, the degree to which the range of outcomes was within the range of sensor accuracy was found to differ for each variable (Table 3: 99.4% for temperature, 60.9% for pH, and 17.9% for specific conductance). The greatest differences in the processed data from different technicians occurred at points of calibration, many of which were outside the ranges of sensor accuracy (Fig. 6). This result is somewhat expected given that the manufacturer-reported accuracies assume a recently calibrated sensor. Because the interval between calibrations in the experiment was typically 2 weeks, it is reasonable to assume that the actual range of sensor accuracy that could be achieved in the field is considerably larger than what is reported by the manufacturer (Thoma et al., 2012; Wagner et al., 2006).

3.4. Comparison of summary statistics

To provide additional context to the variability between users and impact on the resulting datasets, we calculated summary statistics for each variable and report the ranges across all participants (Table 4). For most of these metrics, the ranges for temperature and pH were similar for all statistics and were lower than the manufacturer reported accuracy. However, there are a few exceptions. The range for the 75th percentile of temperature (0.23° C) is outside of the manufacturer's accuracy (± 0.1° C), but still within the USGS criteria for “Excellent” accuracy rating (± 0.3° C), and the ranges for the minimum and maximum of pH exceed both standards. For specific conductance, the ranges are all outside of the manufacturer reported accuracy (± 2 μS/cm) though many are within the USGS “Excellent” criteria (3% or 12 μS/cm).

For pH and specific conductance, the central tendencies exhibit a lower range across participants than do the ranges at the extremes. Artifacts of errors made by participants (described in Section 3.7) are exhibited in the minimum and maximum, but do not affect the central tendencies and percentiles. The ranges of these statistics show that the variability of results between participants may be small enough to be neglected for many purposes, though egregious mistakes have outsized effects. For calculations that are made based on these data, such as mean annual temperature or percent exceedance of water quality criteria, the differences between participants' results are likely not great enough to be meaningful in the assessment of compliance or to add uncertainty in the determination of a summary statistic.

Table 5
Quantification of procedures selected by participants for handling spurious data. Percentages are averages of the proportion of data that was handled by either deletion or setting to −9999 across the participants in each group.

Group	Procedure	Temperature	pH	SpCond
Experienced	−9999	1.62% (n = 13)	7.35% (n = 13)	1.85% (n = 13)
Novice	−9999	1.01% (n = 15)	3.53% (n = 10)	1.15% (n = 3)
Experienced	Deleted	0.14% (n = 13)	0.41% (n = 13)	0.13% (n = 13)
Novice	Deleted	0.00% (n = 15)	0.02% (n = 10)	0.00% (n = 3)

3.5. Handling spurious data

Another aspect for which discrepancies were found among participants' post processed data was in removal of questionable data values. Inconsistency was observed both in which values were deemed to be questionable and subsequently removed and in the procedure used to remove these data. Some participants removed what they considered to be obviously bad data from the corrected record while other participants set problematic values to a "No Data" value (−9999) and then added qualifier flags to note issues with the data (e.g., "sensor malfunction"). Experienced participants set more data to "No Data" and/or deleted more data than did novice participants (Table 5), demonstrating novice participants' reluctance to change data as well as a lack of consistency between implementation of QC practices. In general, participants implemented only one of these procedures for handling spurious data for all series. The guidance explicitly provided to novice participants was to set data to −9999 and add flags; therefore, essentially no participants in this group deleted data (Table 5). Our assumption was that experienced participants would follow this same guidance; however, because we did not remind everyone in this group about this protocol, some experienced participants opted to remove values from the corrected record in lieu of setting data to −9999.

3.6. Details in QC script

In addition to reviewing processed data, the level of detail of annotations that participants included (or did not include) in their QC scripts was examined to document their decisions and actions related to QC. Discrepancies were found in the degree to which participants added commentary to their data editing Python scripts. The level of detail in script comments was rated as Low (little to no comments), Medium (user documented actions without detail), and High (comment details provide insight into decision making). We anticipated that experienced participants would include more verbose and descriptive comments than novice participants; however, the differences between novice and experienced participants was not significant (Fig. 8) according to the chi-square test (Jones et al., 2016). The practice of making comments in scripts was described in the orientation with novice participants, but we did not have high expectations for implementation. However, some novice participants included significant detail in their scripts. Though this practice is a regular and recommended part of our QC workflow, there is obviously variation in how fully participants implement commenting.

Though ancillary to actual processed results, comments in scripts generated by participants can provide insight into QC decisions and are important when scripts are reviewed to trace or reproduce the QC process. The level of detail in comments and annotations in the scripts affects the reproducibility of results. If comments are included, data users and QC technicians may better understand the decision-making process - such as the rationale for selection of a particular offset for drift correction.

3.7. Errors in the QC process

In reviewing results, we noted several participants who made what we consider egregious mistakes in their quality control decisions. In one case (Fig. 9a) a novice user decided to drift correct temperature data. We assume that this choice was made to close the data gap associated with a disconnected sensor during the September 16, 2014 calibration; however, there are no comments in the script to confirm. In this instance, all other users interpolated this short gap and added a qualifier to flag the period. We conclude that this participant either did not understand the scenarios in which drift correction is an appropriate QC procedure (i.e., for sensors that undergo regular calibration) or lacked familiarity with the QC software and mistakenly selected the drift correction button rather than another function.

Results showed that experienced quality control technicians are not immune to mistakes in the QC process. In one case (Fig. 9b), an experienced participant applied an offset of −3 to correct a pH calibration. This was clearly an error, and we speculate that the participant intended to apply a − 0.3 offset to the drift correction given that the average offset applied by other participants for this calibration was −0.23. (Note that a similar error was observed (not shown) for a novice user choosing an unexpectedly large offset while drift correcting pH.) In another instance (not shown), an experienced participant incorrectly set the dates for all linear drift corrections. Instead of setting the starting point of the correction after the previous calibration, this participant set all corrections to begin January 1, 2014. These examples underscore the need for careful training and data review, particularly among those less familiar with QC, but also for experienced technicians. Furthermore, these erroneous decisions may influence the results reported in Section 3.3 – excluding these data would have provided a tighter range of results that might have been more fully within sensor accuracy.

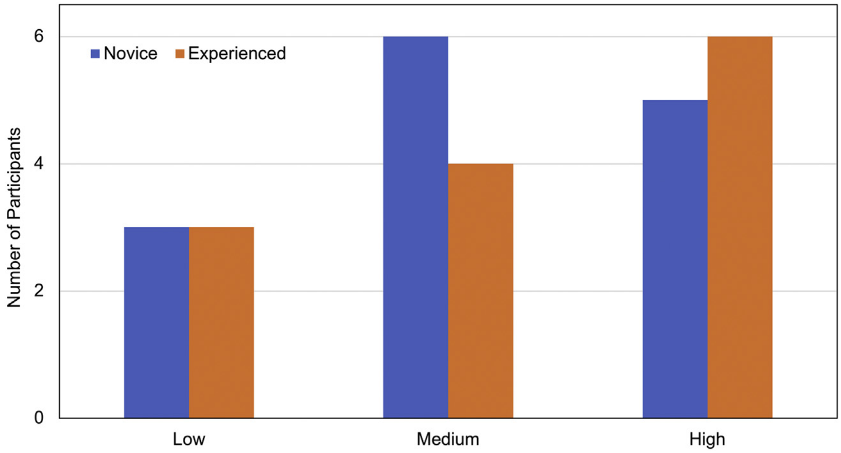

Fig. 8. Determination of level of detail of comments in participants' QC scripts organized by experience level.

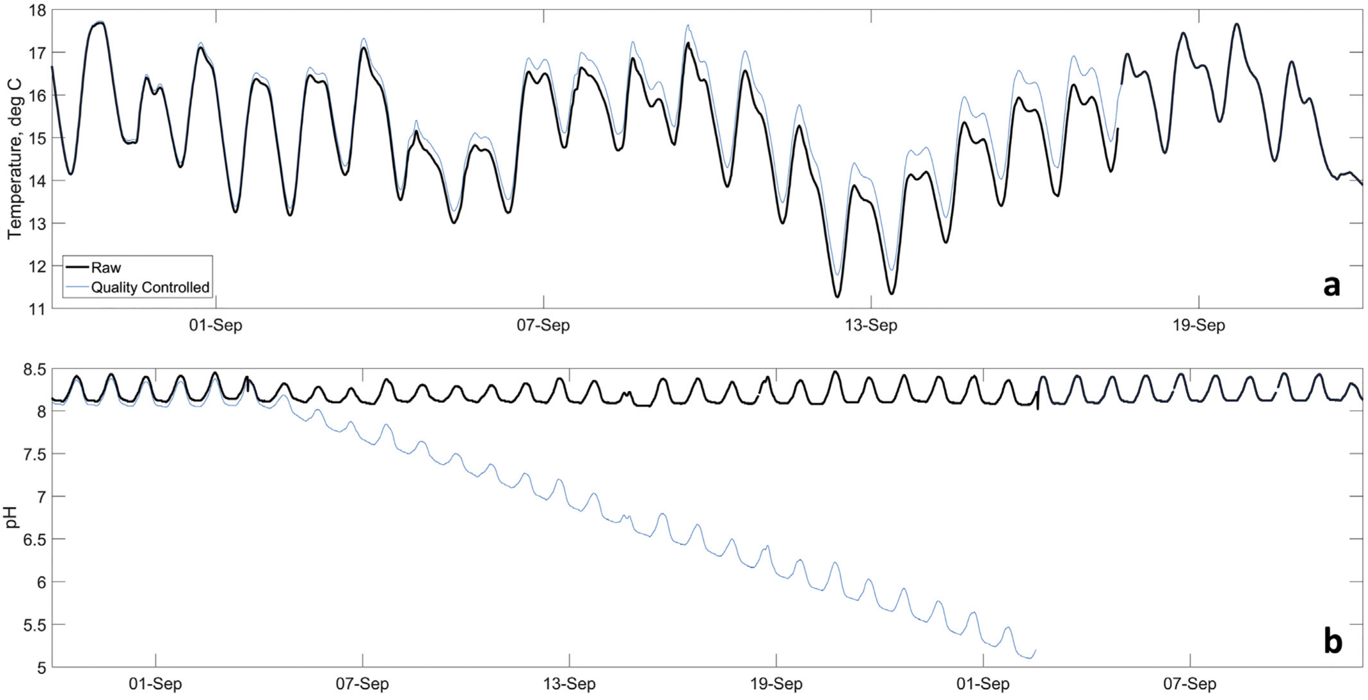

Fig. 9. Examples of QC errors performed by a novice participant on a temperature (a) and an experienced participant on pH (b).

4. Recommendations

This experiment revealed variability in processed datasets resulting from several aspects of QC for which improvements could be made. Though it is likely that QC on these types of data will retain elements of subjectivity, steps can be taken to improve consistency between users. Many discrepancies between users' processed results could be avoided by implementing more specific guidelines that are clearly communicated to all technicians. For example, the differences between handling periods of spurious data could be eliminated by consistently applying QC guidelines. Despite our best efforts to establish these guidelines in our own monitoring network, it is difficult to adequately document the range of decisions that technicians face when performing QC. Therefore, we also recommend that monitoring networks provide ongoing training and discussion for technicians to familiarize and reinforce the application of QC protocols to real data and to bolster consistency in the application of guidelines between practitioners. The short training session provided to novices in this exercise was insufficient, and even some experienced technicians made what we might consider “novice” mistakes.

The greatest periods of discrepancy in processed datasets followed calibration events. To minimize inconsistencies, greater specificity in QC protocols is recommended. For example, data managers and technicians should determine the duration after which the sensor has equilibrated to ambient conditions and valid measurements are being made. This may involve a laboratory experiment, a review of numerous past calibration events, and/or consultation with sensor manufacturers. Implementing a standard post-calibration equilibration period will eliminate inconsistencies in the determination of a valid point after calibration, resulting in less variability in linear drift correction offsets. To further narrow the selection of offsets, more detailed field notes regarding pre and post calibration values and the timing of returning sensors to ambient conditions are recommended. For example, the USGS standard practice is to use pre and post calibration values to set drift correction offsets, though this assumes isolation of drift due to calibration from drift caused by sensor fouling and resolved by sensor cleaning (Wagner et al., 2006).

Variability in different technicians' QC results may not be significant in interpreting finalized data if it is within the range of sensor accuracy. However, it does contribute to the overall uncertainty of the observations. If very accurate measurements are important for a given application, a collaborative process supporting QC decisions is suggested. Results reported here corroborate the recommendations of the USGS, which require review of corrected datasets for “completeness and accuracy” by two professionals in addition to the original hydrographer (Wagner et al., 2006). Indeed, in other domains, thorough reviews and a team-based process have been recommended to improve the quality of assessments (Banghart et al., 2016), and a distributed and federated approach was found to improve consistency of modeling results (Stockhause et al., 2012). In one case within hydrology, Neal et al. (2013) report collaborative data processing based on multiple, expert opinions as “lively and sustained discussion.” Multiple technicians reviewing data could lead to developing consensus about each instance of a QC procedure, which would reduce subjectivity and result in more consistent results. In the current QC workflow for the GAMUT network, QC work and scripts by less experienced technicians are reviewed and data are saved by the most experienced technician in each watershed; however, not all processed data is reviewed consistently, and most data are not reviewed at the level of granularity to consider a range of possible offsets for linear drift corrections. Data review also helps ensure that egregious mistakes are identified, avoiding situations like those described in Section 3.6.

Although QC evaluation by multiple technicians at a granular level could reduce the variability observed in this study, it is acknowledged that this level of review may not be feasible for all applications, and it may even be more detailed than what is recommended by the USGS. Scientists aim to produce the best data possible, but the potential improvement in results may not be significant enough to warrant the time and resources that collaborative QC and detailed review would require. Depending on the required level of data quality, performance of fine scale QC by a trained technician under specific QC guidelines followed by review by another trained technician to assess that data meet general guidelines should generate processed datasets that meet the needs for most scientific studies. In lieu of review by two separate technicians, a single technician might perform QC and then re-review data in the full context of the raw data as well as data from other sensors and other monitoring stations.

5. Conclusions

This study examined the results of individual participants performing QC on identical datasets of raw, high frequency water quality data using consistent tools and guidelines. Two groups comprised of novice and experienced participants were included to examine whether results differed based on fluency with these types of data and with QC procedures. Despite expectations that experienced participants would produce more consistent results and that novice participants would make more errors in performing QC, there was greater variability in experienced participants' processed results than those of novices. We conclude that novices' unfamiliarity with QC procedures resulted in hesitancy to alter data.

The periods of greatest discrepancy followed field calibration events that necessitated drift corrections in the QC process. As a result, there was little variability in the processed datasets of variables that do not undergo calibrations. We found that, depending on the observed variable, the variability was within the range of sensor accuracy, but for those periods associated with calibration events, the discrepancies resulting from the QC process exceeded sensor accuracy to a varying degree. To improve consistency, clarifying QC guidelines and protocols and thoroughly training technicians is recommended. Implementing a collaborative QC process is also suggested wherein the changes introduced by QC for sensitive periods are reviewed for cases where highly accurate data are required. Because of the resources demanded by review and collaboration, in determining QC workflows, scientists should look to balance the level of review with the potential improvements in processed data quality and precision.

Acknowledgments

This work was supported by the U.S. National Science Foundation under EPSCoR grant IIA-1208732 awarded to Utah State University as part of the State of Utah EPSCoR Research Infrastructure Improvement Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors gratefully acknowledge the contributions of the participants who took part in this study.

Declarations of interest

None.

References

Banghart, M., Babski-Reeves, K., Bian, L., 2016. Human induced variability during failure mode effects analysis. In: Proceedings - Annual Reliability and Maintainability Symposium 2016–April, http://dx.doi.org/10.1109/RAMS.2016.7448000.

Campbell, J.L., Rustad, L.E., Porter, J.H., Taylor, J.R., Dereszynski, E.W., Shanley, J.B., Gries, C., Henshaw, D.L., Martin, M.E., Sheldon, M., Wade, Boose, E.R., 2013. Quantity is nothing without quality. Bioscience 63, 574–585. http://dx.doi.org/10.1525/bio.2013.63.7.10.

Daly, C., Redmond, K., Gibson, W., Doggett, M., Smith, J., Taylor, G., Pasteris, P., Johnson, G., 2005. Opportunities for improvements in the quality control of climate observations. In: 15th AMS Conference on Applied Climatology, pp. 20–23.

Dereszynski, E.W., Dietterich, T.G., 2007. Probabilistic models for anomaly detection in remote sensor data streams. In: Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI2007), pp. 75–82.

ESIP EnviroSensing Cluster, 2014. Community wiki document on best practices for sensor networks and sensor data management. In: Federation of Earth Science Information Partners, . http://wiki.esipfed.org/index.php/EnviroSensing_Cluster, Accessed date: 1 January 2016.

Fiebrich, C.A., Morgan, C.R., McCombs, A.G., Hall, P.K., McPherson, R.A., 2010. Quality assurance procedures for mesoscale meteorological data. J. Atmos. Ocean. Technol. 27, 1565–1582. http://dx.doi.org/10.1175/2010JTECHA1433.1.

Gries, C., Henshaw, D., Brown, R.F., Cary, R., Downing, J., Jones, C., Kennedy, A., Laney, C.M., Martin, M., Morse, J., Porter, J., Read, J.S., Rettig, A., Sheldon, W., Strachan, S., Zdravkovic, B., 2014. Sensor and sensor data management best practices released. In: LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network, Spring.

Hart, J.K., Martinez, K., 2006. Environmental sensor networks: a revolution in the earth system science? Earth Sci. Rev. 78, 177–191. http://dx.doi.org/10.1016/j.earscirev.2006.05.001.

Hill, D.J., Minsker, B.S., Amir, E., 2009. Real-time Bayesian anomaly detection in streaming environmental data. Water Resour. Res. 45, 1–16. http://dx.doi.org/10.1029/2008WR006956.

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I., 2008. A relational model for environmental and water resources data. Water Resour. Res. 44. http://dx.doi.org/10.1029/2007wr006392.

Horsburgh, J.S., Reeder, S.L., Jones, A.S., Meline, J., 2015. Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. Environ. Model. Softw. 70. http://dx.doi.org/10.1016/j.envsoft.2015.04.002.

Jones, A.S., Horsburgh, J.S., Jackson-Smith, D., Ramírez, M., Flint, C.G., Caraballo, J., 2016. A web-based, Interactive Visualization Tool for Social Environmental Survey Data. Environ. Model. Softw. 84. http://dx.doi.org/10.1016/j.envsoft.2016.07.013.

Jones, A.S., Aanderud, Z.T., Horsburgh, J.S., Eiriksson, D.P., Dastrup, D., Cox, C., Jones, S.B., Bowling, D.R., Carlisle, J., Carling, G.T., Baker, M.A., 2017. Designing and implementing a network for sensing water quality and hydrology across mountain to urban transitions. J. Am. Water Resour. Assoc. http://dx.doi.org/10.1111/1752-1688.12557.

Jones, A.S., Eiriksson, D.P., Horsburgh, J.S., 2018. Quality Control Experiment. HydroShare. http://dx.doi.org/10.4211/hs.31f30d14c88748d986842d278d125a5c.

Meek, D.W., Hatfield, J.L., 1994. Data quality checking for Single Station meteorological databases. Agric. For. Meteorol. 1923, 25.

Moatar, F., Miquel, J., Poirel, A., 2001. A quality-control method for physical and chemical monitoring data. Application to dissolved oxygen levels in the river Loire (France). J. Hydrol. 252, 25–36. http://dx.doi.org/10.1016/S0022-1694(01)00439-5.

Neal, C., Reynolds, B., Kirchner, J.W., Rowland, P., Norris, D., Sleep, D., Lawlor, A., Woods, C., Thacker, S., Guyatt, H., Vincent, C., Lehto, K., Grant, S., Williams, J., Neal, M., Wickham, H., Harman, S., Armstrong, L., 2013. High-frequency precipitation and stream water quality time series from Plynlimon, Wales: an openly accessible data resource spanning the periodic table. Hydrol. Process. http://dx.doi.org/10.1002/hyp.9814.

Pastorello, G., Agarwal, D., Papale, D., Samak, T., Trotta, C., Ribeca, A., Poindexter, C., Faybishenko, B., Gunter, D., Hollowgrass, R., Canfora, E., 2014. Observational data patterns for time series data quality assessment. In: Proceedings - 2014 IEEE 10th International Conference on eScience, eScience 2014. vol. 1. pp. 271–278.

Pellerin, B.A., Stauffer, B.A., Young, D.A., Sullivan, D.J., Bricker, S.B., Walbridge, M.R., Clyde, G.A., Shaw, D.M., 2016. Emerging tools for continuous nutrient monitoring networks: sensors advancing science and water resources protection. JAWRA J. Am. Water Resour. Assoc. 20460, 1–16. http://dx.doi.org/10.1111/1752-1688.12386.

Qu, W., Bogena, H.R., Huisman, J.A., Schmidt, M., Kunkel, R., Weuthen, A., Schiedung, H., Schilling, B., Sorg, J., Vereecken, H., 2016. The integrated water balance and soil data set of the Rollesbroich hydrological observatory. Earth Syst. Sci. Data 8, 517–529. http://dx.doi.org/10.5194/essd-8-517-2016.

Rode, M., Wade, A.J., Cohen, M.J., Hensley, R.T., Bowes, M.J., Kirchner, J.W., Arhonditsis, G.B., Jordan, P., Kronvang, B., Halliday, S.J., Skeffington, R.A., Rozemeijer, J.C., Aubert, A.H., Rinke, K., Jomaa, S., 2016. Sensors in the stream : the high-frequency wave of the present. Environ. Sci. Technol. http://dx.doi.org/10.1021/acs.est.6b02155.

Shafer, M.A., Fiebrich, C.A., Arndt, D.S., Fredrickson, S.E., Hughes, T.W., 2000. Quality assurance procedures in the Oklahoma Mesonetwork. J. Atmos. Ocean. Technol. 17, 474–494. http://dx.doi.org/10.1175/1520-0426(2000)017<0474:QAPITO>2.0.CO;2.

Sheldon, W.M., 2008. Dynamic, rule-based quality control framework for real-time sensor data. In: Gries, C., Jones, M.B. (Eds.), Proceedings of the Environmental Information Management Conference. Albuquerque, NM, pp. 145–150.

Stockhause, M., Höck, H., Toussaint, F., Lautenschlager, M., 2012. Quality assessment concept of the world data Center for Climate and its Application to CMIP5 data. Geosci. Model Dev. 5, 1023–1032. http://dx.doi.org/10.5194/gmd-5-1023-2012.

Taylor, J.R., Loescher, H.L., 2013. Automated quality control methods for sensor data: a novel observatory approach. Biogeosciences 10, 4957–4971. http://dx.doi.org/10.5194/bg-10-4957-2013.

Thoma, D.P., Irwin, R.J., Penoyer, P.E., 2012. Documenting measurement sensitivity and Bias of field-measured parameters in water quality monitoring programs. Environ. Monit. Assess. 184, 5387–5398. http://dx.doi.org/10.1007/s10661-011-2347-5.

Wagner, R.J., Boulger, R.W., Oblinger, C.J., Smith, B.A., 2006. Guidelines and Standard Procedures for Continuous Water-Quality Monitors: Station Operation, Record Computation, and Data Reporting. http://pubs.usgs.gov/tm/2006/tm1D3/.

White, D.L., Sharp, J.L., Eidson, G., Parab, S., Ali, F., Esswein, S., 2010. Real-time Quality Control (QC) processing, notification, and visualization services, supporting data management of the intelligent river. In: Proceedings of the 2010 South Carolina Water Resources Conference, pp. 4.

Xylem, 2012. EXO User Manual, Item# 603789REF, Revision. F. YSI Incorporated, Yellow Springs, OH. https://www.ysi.com/File Library/Documents/Manuals/EXO-User-Manual-Web.pdf.